Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 16:48:42
REFERENCE NO: 295

Author Names & Affiliations

- Eric Michielssen - Louise Ganiard Professor of Engineering, Professor of Electrical and Computer Engineering, Associate Vice President, Advanced Research Computing
- Sharon Broude Geva - Director of Advanced Research Computing

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Various (see attributions in text)

Title of Submission

University of Michigan response to NSF RFI on Future Needs for Advanced Cyberinfrastructure

Abstract (maximum ~200 words).

Recent developments in Computational and Data Science and Engineering (CDSE) have led to a heightened reliance on Advanced CyberInfrastructure (ACI) by researchers from diverse backgrounds, thereby intensifying the need for sophisticated and integrated hardware, software, and network capabilities, expanded computing capacity at all resource tiers, new capabilities for interaction and gateway access to resources, and an increased focus on providing and developing the human component of the ACI ecosystem.

The NSF should actively support the needs of CDSE researchers. We expect consumption of ACI resources by University of Michigan (U-M) researchers to expand significantly over the next 5 years. Growth is expected across all resource tiers and usage modalities but especially in the use of public and private cloud resources that enable data-intensive discovery and engineering. U-M researchers would benefit from programs that support the development of software and associated data sources, along with their maintenance, distribution, and publication. U-M prides itself on supporting highly interdisciplinary research initiatives, which undoubtedly would benefit from enhanced collaboration between the NSF and NIH through joint funding programs, development of ACI resources, and guidelines and tools for research data management.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

U-M research in CDSE is highly interdisciplinary in nature and supported through active research programs in the university’s 19 schools and colleges as well as large-scale centrally funded initiatives (computational and data science (www.arc.umich.edu), mobility and transportation, exercise and sports science, etc).

In preparation of U-M's response to this RFI, CDSE faculty and researchers from across the institution were asked to comment on emerging science and engineering research challenges particular to their field of inquiry. Representative quotes from faculty are included below. Computational hurdles present in many emerging areas involve (a) requirements for executing truly data science- and HPC-concurrent workflows, (b) new modalities for supporting large-scale data mining, machine learning, and optimization, (c) needs for new techniques for collaborating and sharing information.

Astronomy. "Large investments by NSF are being made to build and execute sky surveys across the electromagnetic spectrum, particularly radio-mm and optical-IR, in order to address the nature of dark energy, dark matter and gravity on cosmic scales. In order to exploit all the information contained in the data, particularly gravitational lensing signatures and clusters of galaxies, a key challenge is to model the universe on all relevant scales, including those of millions of parsecs or below at which the dynamics of the multiple cosmic fluid components are strongly non-linear. Hydrodynamic simulations of cosmic structure formation provide the means to probe the nonlinear evolution of structure in detail, and direct sky survey realizations derived from such simulations offer important pathways to estimate error covariance and to identify and mitigate key sources of systematic error in astronomical survey analysis." (August Evrard, Professor of Physics and Astronomy; Christopher Miller, Professor of Astronomy and Physics).

Material Science, Combustion, Computational Fluid Dynamics. (a) "We are developing a broad new paradigm in computing for rare, but disruptive events. This could include applications such as climate and weather science, ecosystems, energetic materials (e.g., catastrophic failure of batteries), catastrophic failure of materials and structure, combustion, and many more. The challenge will be to combine data, computational science, probabilistic methods, and finally, a public advocacy role. The last item is critical to convey specific, actionable information and recommendations to the public and policy makers." (Krishna Garikipati, Professor of Mechanical Engineering, and Mathematics, Director, Michigan Institute for Computational Discovery & Engineering). (b) Accurate predictions of turbulent multiphase flows are crucial to many engineering and environmental applications. Advancements in numerical modeling and growing computational resources are beginning to enable quality predictions of turbulent multiphase flows. Predictive capabilities, however, do not by themselves lead to new insight or design improvement. The formulation of robust design rules requires careful interpretation of simulation results, and for design optimization, requires repeated predictive runs. For large-scale simulations, the volume and rate at which data is produced present significant challenges in obtaining the necessary insight to improve design. Manual analysis of large-scale heterogeneous datasets is becoming increasingly impractical due to these big data challenges. With exascale computing in the foreseeable future, new analysis tools will be essential in providing physical insight from predictive simulations. Moreover, the amount of data required to perform apriori analyses will not be feasible at exascale. New tools that are (a) non-intrusive and (b) portable on emerging heterogeneous architectures will need to be developed." (Jesse Capecelatro, Professor of Mechanical Engineering)

Large-Scale Data Mining, Machine Learning and Optimization for Energy Systems, Transportation, and Computational Social Science. "In energy systems, it is necessary to reconsider the main assumptions underlying the grid and to solve the resulting machine learning and optimization problems. These are substantial computational challenges given the uncertainties and contingencies to take into account, and the nonlinearities of energy systems. In transportation, the computational challenges range from large-scale planning problems to optimization and machine learning for real-time operations. Sensing Infrastructure for Transportation to create high-fidelity spatial, temporal, and contextual mobility patterns. Transportation needs a significantly upgraded sensing architecture. The US is well behind countries like Australia, which are much smaller." (Pascal Van Hentenryck, Professor of Industrial and Operations Engineering, and Computer Science and Engineering)

Weather and climate modeling. "Flows in both the atmosphere and ocean take place over a vast array of time and space scales. This means that the CPU and storage demands for atmospheric, oceanic, and climate modeling are immense." (Brian Arbic, Professor of Earth and Environmental Sciences, and Climate and Space Science)

Modeling of Risk. "The modeling analysis of risk is becoming a key topic in several areas, including estimating low probability events with high financial and societal impact. Examples include predicting the risk of failure of a new aircraft engine, the probability of genetic mutations that dictate evolution, the physics that lead to extreme weather patterns etc. More recently, such tools have been extended to non-physics areas such as financial economics." (Venkat Raman, Professor of Aerospace Engineering)

Precision Health. "Precision health refers to the use of population-based strategies aimed at discovering and validating markers that influence disease prevention and health outcomes, and that can subsequently be used to make actionable decisions to personalize an individual's pursuit of wellness. The field requires the analysis and integration of numerous data sources (genomic, images, electronic health record, social, etc.) in large-scale secure environments tightly integrated with massive storage resources using modern data science

/ database / artificial intelligence methods.” (Many faculty)

Scientific Interaction and Collaboration. “Science is increasingly a team sport. We need to provide infrastructures that recognize, support and encourage the increasingly distributed, collaborative nature of science while striving to hide the complexity of that infrastructure from scientists. Scientists need to be able to focus on their research without having to become infrastructure experts. However, at least some scientists and some infrastructure providers need to interact with each other to ensure we develop systems, architectures, and facilities that are effective in supporting science. Entities like Open Science Grid can help centralize and sustain expertise and connect multiple institutions, filling gaps in knowledge, capability, and expertise that individual institutions typically cannot provide.” (Shawn McKee, Research Scientist, Physics and Advanced Research Computing (ARC))

Learning Analytics. “Ongoing research into how learning occurs has the potential to revolutionize the field of education. However, investigations into learning analytics are often limited by the ability to get large, accurate datasets about patterns of interaction between students and between students and faculty, relying instead on voluntary participation by unreliable narrators.” (Eric Boyd, Director of Research Networks, Information and Technology Services)

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

U-M researchers extensively use all tiers of computing resources ranging from individual/lab-based systems to campus, national, and leadership infrastructure. Future research will benefit from an expansion of present compute, storage, network, and software resources across all tiers. Additionally, future research will require new systems with technical capabilities that go beyond those of today’s resources. Some of these capabilities likely will be derived from experimental systems being developed as part of NSF-funded projects (such as the ConFlux MRI and OSiRIS DIBBS at U-M), provided they can be scaled up and deployed as national scale resources or replicated in other environments, e.g. public clouds. The needed technical advancements include, but are not limited to (i) compute systems for data-intensive simulations; (ii) infrastructure designed with monitoring in mind; (iii) programmable networks integrated into infrastructure; (iv) infrastructure designed for collaborative, distributed use; (v) energy-conscious applications to enable use of the next generation supercomputers; (vi) computational frameworks that allow the linking of computational models from different fields into one integrated environment; (vii) new infrastructure elements to support persistent data collections and researcher identity across multiple collaborations. Representative quotes from faculty and researchers are included below:

On software, data curation, and downstream data processing. “NSF support of computational cosmology has historically been fragmented and incoherent. As a result, multiple groups develop quasi-independent code bases featuring redundant physics. Most of these codes remain proprietary, and thus cannot be tested and developed through open source community efforts. NSF should take seriously the call of the 2010 Decadal Survey of Astronomy and renew support for Theoretical and Computational Astrophysics Networks (TCAN) with a special focus dedicated to open source computational cosmology. This renewed commitment would be particularly powerful if coordinated with XSEDE resource allocation. Existing cyberinfrastructure is largely sufficient for the task, but curation of data collections and organization of downstream data processing for the astrophysical and cosmological communities would require new infrastructure elements to support persistent data collections and researcher identity across multiple collaborations.” (Evrard, Miller)

On data-intensive computing. “There still aren't large- scale clusters that can handle data processing and high performance computing for physics simultaneously. At U-M, we have made a start with the ConFlux cluster that executes data access, machine learning and scientific computing for physics simultaneously.” (Garikipati)

On the cost of computing and availability of cycles. “The costs of computing for my group is likely to reach about $1 million a year in coming years as my students progress in their PhDs. There is a strong need to consider HPC as a critical service for faculty in engineering given the computational demands of the next generation of software for addressing fundamental societal issues.” (Van Hentenryck). (b) “As with other scientific communities, the atmospheric/oceanic/climate modeling communities could benefit greatly from NSF systems that deliver more CPU cycles and greater storage capability. The ability to move model output quickly between institutions, thus allowing more scientists to analyze large model outputs, would also greatly benefit our community.” (Arbic)

On system monitoring, importance of networks, and seamless collaboration. “I see three primary limitations with existing cyberinfrastructure: 1) It is too opaque; there is not enough visibility into what it is happening at any point in time. Monitoring that exists is either too specific, or not easily integrated to provide a clear picture of how the system is performing at a high-level. We need infrastructure components designed with monitoring as a primary consideration from the start. We need systems designed to incorporate lower- level monitoring and augment it with system- level monitoring to better understand, maintain and optimize our complex infrastructures. 2) Networks will be increasingly important as science moves beyond the building or cluster. Networks must be able to be integrated as seamlessly as possible into our evolving infrastructures. There are significant challenges in how best to do this across two or more administrative domains. How can science domains be given part of the network for their use? What mechanisms can support sharing programmable networks between many science domains? Ultimately, networks are a foundational component of cyberinfrastructure just like compute, storage, middleware and people. 3) Infrastructures must be designed with collaborative, distributed users in mind from the start. As science collaborations grow and evolve, the infrastructure supporting them must be capable of transparently growing and evolving as well. Storage needs to useable across collaborations, not just local to a cluster. Identity and authorization need to be built from institutional identities (not creating new users/passwords for each new activity). Systems must be secure but still able to easily support distributed, multi-institutional collaborations. Software (middleware and applications) needs to be designed up front to support this collaborative, distributed science team paradigm.” (McKee)
“Today’s wireless access points are by design “dumb terminals” designed solely to engender connectivity. If we can transform today’s wireless access points into a next-generation mesh of edge node programmable sensors of human activity, researchers would have the ability to ask all manner of questions about how people move, communicate, acquire information, and make decisions. The potential to reveal new insights into aggregate human behavior is immense. Assuming useful data can be gleaned from such location-based operational data, the ability to exploit such data while preserving privacy can be challenging even with modern statistical techniques in differential privacy. New research is required into how best to find patterns in the movements of autonomous actors that do not reveal the identity of individual persons.“ (Boyd)

On the needs of interdisciplinary research. “Each of the world’s engineering grand challenges is being tackled by a multitude of highly specialized sub-disciplines, each dedicated to handling a subset of the overall challenge. Although it appears logical that computational researchers in various fields should be able to collaborate with minimal effort because they speak the same language,” the reality is that the disciplinary boundaries remain high because the vast majority of computational models were not designed from the start to be compatible with one another. Research is needed to develop: 1) high- level testbed problems that realistically represent each grand engineering challenge being considered, 2) standardized, robust, web-accessible, plug-and-play, and easy-to-use computational frameworks that allows researchers from different sub-fields to link their computational models together to study the effects of each discipline on the overall problem within an integrated environment, 3) standardized validation methodologies to ensure that the developed simulation models yield reasonable results, and 4) standardized visualization techniques that allow the results to be conveniently viewed and understood.” (Sherif El-Tawil, Professor of Civil & Environmental Engineering, and Associate Chair)

On needs for precision health research. Precision health presents the challenge of bringing together data types from medicine, social and behavioral, and traditional sciences to gain new insights in, and build actionable predictive models that enhance the health of individuals and entire populations. Each data source falls under one of many regulatory or privacy domains that present shared infrastructure cannot support. ACI that supports software reconfigurable enclaves are called for that allow feature customization (big compute, big data, machine learning) and data isolation between projects, in a nimble and cost effective way, avoiding significant overhead involved with lab level (compliant and customized), and shared (cost effective and scalable) computational resources. (Brock Palen, Director, Advanced Research Computing - Technology Services)

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

We fully support and endorse the CASC response to this RFI.
In addition, U-M faculty and researchers offer the considerations below that are not covered in questions 1 and 2.

On workforce development. Although CI, broadly defined, includes software, hardware, and networks as well as people, it is important to note that the human aspects are often the most difficult to fulfill. There is need for learning and workforce development for a broad range of CI-related career paths, including computational scientists for industry research and production and for academic research, HPC operators, research software developers, and domain specialists who can serve as consultants to students and faculty. (Many faculty)

“It is critical that we find ways to encourage and support “Cyber-practitioners” who bridge between science domains and cyberinfrastructure. How to do this sustainably across multiple institutions? There have to be ways to recognize and support those filling this role and recognition and support need to come both from academia and the local institutional leadership. It is not feasible to develop, support and extend the evolving infrastructures we discuss in questions 1 and 2 without creating a professional, intellectual corp of such practitioners.“ (McKee)

On software. Sustainability of software created as part of research projects, and retention and broad access to research data are great concerns for the research community at U-M. These require easy-access paradigms to incentivize re-use rather than duplication, and funding for medium- and long-term solutions to serve research beyond the boundaries of individual campuses. (Many faculty)

On organizational matters. Related to the organization and process aspects are coordination and administration tasks and requirements that can become barriers to utilization of NSF awarded resources. Two examples are:
(a) awarded access to compute and storage resources (including expertise for their usage) are not always interlinked with funding for research projects, creating additional obstacles. “While XSEDE supercomputing resources offer world-class computational platforms to produce and analyze cosmological simulations, a structural deficit in these allocations is the fact that they are decoupled from funding allocations made by AAG and other NSF divisions. As a result, PIs of computational cosmology efforts face a double review process that reduces the overall likelihood of success (the “epsilon-squared” problem) and that also can lead to perverse outcomes, such as personnel funding with no cycles or cycles with no money for personnel. This inefficiency ultimately limits the growth potential of computational cosmology. In particular, group efforts are strongly discouraged - effectively, if not literally - under the current, sole-PI funding climate.” (August Evrard, Christopher Miller)
(b) Without NSF certification of awarded commercial cloud resources, researchers from multi-institution projects could face significant risks and/or administrative overhead working collaboratively (e.g., in cases where each institution has either a separate BAA with the cloud provider or not all the institutions have one at all).

On networking funding and development. “The rise of data-intensive science is creating an ever-widening gap between the networking needs of the few (data-driven researchers) and the many (consumers and enterprise workers). The networking needs of the former group are growing exponentially in comparison to the networking needs of the latter, whose demands are largely bound by the maximum networking needs of mobile platforms connecting to very high definition video. This emerging bandwidth gap raises the question of the long-term viability of network funding models, as the bulk of the network growth is coming from the first, small group of users, yet the funding of such network growth is assumed to come from the second, large group of users. It is incumbent on the NSF to fund not just the building of advanced networks (through CC* and other funding programs) but to develop a strategy for supporting the ongoing operations of such advanced networks whose scale is designed primarily to support the needs of big data researchers. The advent of software-defined networking heralded the ability to decouple network infrastructure into modular components such as the data plane and the control plane. Such modularity has the promise to spur innovation by allowing individual researchers to demonstrate new methods of engendering network connectivity. However, unlike pure software, where modularity through microservices-based-design and containerization is readily achievable, networking has a strong hardware component that allows manufacturers to create closed ecosystems that resist modularization and impede innovation. Over time, such closed systems are likely to impede the ability of network researchers to experiment with and demonstrate new networking ideas. The NSF should support the development of open hardware platforms for network innovation that facilitate network research and experimentation.” (Boyd)

On computational medicine, precision health, bioinformatics. U-M prides itself on supporting highly interdisciplinary research initiatives, which undoubtedly would benefit from enhanced collaboration between the NSF and NIH through joint funding programs (similar to current cross-directorate programs), development of ACI resources and guidelines and tools for research data management. (Many faculty)